Functional Partial Least-Squares: Optimal Rates and Adaptation

Andrii Babii*[1], Marine Carrasco†[2] and Idriss Tsafack‡[2]

[1]Department of Economics, UNC-Chapel Hill

[2]Department of Economics, University of Montreal

February 16, 2024

Abstract

We consider the functional linear regression model with a scalar response and a Hilbert space-valued predictor, a well-known ill-posed inverse problem. We propose a new formulation of the functional partial least-squares (PLS) estimator related to the conjugate gradient method. We shall show that the estimator achieves the (nearly) optimal convergence rate on a class of ellipsoids and we introduce an early stopping rule which adapts to the unknown degree of ill-posedness. Some theoretical and simulation comparison between the estimator and the principal component regression estimator is provided.

Keywords: Adaptive Estimation, Conjugate Gradient Method, Convergence Rates, Functional Linear Regression, Functional Partial Least Squares.

JEL classification: C38, C51, C53, C55, C58.

*Department of Economics, University of North Carolina at Chapel Hill — Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: babii.andrii@gmail.com

†Department of Economics, University of Montreal. Email: marine.carrasco@umontreal.ca. The author thanks NSERC for partial financial support.

‡Department of Economics, University of Montreal. Email: idriss.tsafack.teufack@umontreal.ca

1 Introduction

1.1 Functional Linear Regression

Let $(Y, X) \in \mathbb{R} \times \mathbb{H}$, where $(\mathbb{H}, \langle ., . \rangle)$ is a separable Hilbert space with the induced norm $\|.\| = \sqrt{\langle ., . \rangle}$. The functional linear regression model is

$$Y = \langle \beta, X \rangle + \varepsilon, \qquad \mathbf{E}[\varepsilon X] = 0,$$

where $\beta \in \mathbb{H}$ is the unknown functional slope parameter and for simplicity we assume that $\mathbb{E}[X] = 0$; see Cai & Hall (2006); Cai & Yuan (2012); Cardot et al. (1999, 2003); Cardot & Johannes (2010); Comte & Johannes (2012); Crambes et al. (2009); Delaigle & Hall (2012); Hall & Horowitz (2007) among many other important contributions.

The covariance restriction $\mathbf{E}[\varepsilon X] = 0$ implies that the slope coefficient $\beta \in \mathbb{H}$ solves the moment condition

$$r := \mathbf{E}[YX] = \mathbf{E}[(X \otimes X)\beta] =: K\beta, \tag{1}$$

where $r \in \mathbb{H}$ and $K : \mathbb{H} \to \mathbb{H}$ is a compact covariance operator with summable eigenvalues whenever $\mathbf{E}\|X\|^2 < \infty$. It is well-known that the inverse operator K^{-1} is discontinuous and solving the equation $K\beta = r$ for β is an ill-posed inverse problem; see Carrasco et al. (2007); Engl et al. (1996); Hoffmann & Reiss (2008); Klemelä & Mammen (2010); Nemirovski (1986).

Roughly speaking, there are two popular strategies to regularize such problems:

(a) replace K^{-1} with a continuous operator $R_\alpha(K)$ for some function $R_\alpha : [0, \infty) \to \mathbb{R}$ satisfying $\lim_{\alpha \to 0^+} R_\alpha(\lambda) = \lambda^{-1}$.

(b) solve the problem in a finite-dimensional subspace $\mathbb{H}_m \subset \mathbb{H}$, spanned by some basis vectors $h_1, h_2, \dots, h_m \in \mathbb{H}$.

Examples of (a) include the Tikhonov regularization when $R_\alpha(\lambda) = (\alpha + \lambda)^{-1}$, the spectral cut-off when $R_\alpha(\lambda) = \lambda^{-1}\mathbf{1}_{\lambda \geq \alpha}$ and the Landweber iterations; see Carrasco et al. (2007); Cavalier (2011); Engl et al. (1996) for more details. On the other hand, the estimators in group (b), often solve the empirical least-squares problem

$$\min_{b\in\mathbb{H}_m} \|\mathbf{y} - T_n b\|_n^2, \tag{2}$$

where $\|v\|_n^2 = v^\top v/n, v \in \mathbb{R}^n$ and we put $\mathbf{y} = (Y_1, \dots, Y_n)^\top$ and

$$\begin{aligned} T_n : \mathbb{H} &\to \mathbb{R}^n \\ b &\mapsto (\langle X_1, b\rangle, \dots, \langle X_n, b\rangle)^\top \end{aligned}$$

for an i.i.d. sample $(Y_i, X_i)_{i=1}^n$. The basis $(h_j)_{j=1}^m$ spanning $\mathbb{H}_m$ can be either fixed (e.g. Fourier, polynomials, splines, wavelets) or adaptively constructed from the data.

The data-driven bases are especially attractive since they can adapt to the features of the population represented by the data and can approximate the slope parameter $\beta \in \mathbb{H}$ more efficiently; see Delaigle & Hall (2012). The principal component analysis (PCA)[1] and the partial least-squares (PLS) are two widely used methods to construct adaptive bases in practice. The PCA basis is constructed by identifying the directions in $\mathbb{H}$ where X varies the most while the PLS basis is constructed in a supervised way taking into account the response variable as well. While the first m elements of the PCA basis $h_1, \dots, h_m$ usually capture most of the variation of X, these are not necessarily the most important vectors for approximating β or predicting the response variable Y. It is easy to find empirical examples, where some of the few last low-variance components *are* important; see Jolliffe (1982) who documented the issue on datasets used in economics, climate science, chemical engineering, and meteorology.

1.2 PLS estimator

The PLS estimator constructs a data-driven basis iteratively maximizing the covariance with the response variable Y; see Blazere et al. (2014); Delaigle & Hall (2012); Preda & Saporta

[1] Using the PCA basis is also related to the spectral cut-off method described in (a).

(2005); Reiss & Ogden (2007); Wold et al. (1984) for theoretical analysis and Carrasco & Rossi (2016); Kelly & Pruitt (2015) for applications in economics and finance. The iterative nature of the estimator makes it difficult to analyze its statistical properties. This prompted Delaigle & Hall (2012) to formulate an alternative functional PLS solving the problem in equation (2) over the so-called Krylov subspace

$$\mathbb{H}_m = \operatorname{span}\left\{\hat{r}, \hat{K}\hat{r}, \hat{K}^2\hat{r}, \dots, \hat{K}^{m-1}\hat{r}\right\},$$

where

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i \qquad \text{and} \qquad \hat{K} = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i$$

are the estimators of r and K; see also Helland (1988); Phatak & de Hoog (2002) for the link between PLS and Krylov subspaces.

While the functional PLS estimator is popular in practice due to its efficient representation of the data, to the best of our knowledge, it is still unknown whether it is the minimax optimal estimator. In this paper, we study a variation of the PLS estimator with $m \geq 1$ components, denoted $\hat{\beta}_m$, characterized as a solution to the least-squares problem

$$\min_{b \in \mathbb{H}_m} \|T_n^*(\mathbf{y} - T_n b)\|^2$$

over the Krylov subspace $\mathbb{H}_m$. The least-squares objective function is weighted by the adjoint operator of T_n

$$\begin{aligned} T_n^* &: \mathbb{R}^n \to \mathbb{H} \\ \phi &\mapsto \frac{1}{n}\sum_{i=1}^{n} X_i \phi_i \end{aligned}$$

and corresponds to minimizing the first-order conditions to the problem in equation (2), often called the normal equations. Equivalently, $\hat{\beta}_m$ fits the empirical counterpart to the equation (1)

$$\min_{b \in \mathbb{H}_m} \left\|\hat{r} - \hat{K}b\right\|^2 \tag{3}$$

as it is easy to see that $\hat{r} = T_n^*\mathbf{y}$ and $\hat{K} = T_n^* T_n$. Importantly, the PLS estimator formalized in equation (3) corresponds to the conjugate gradient method with a self-adjoint operator $\hat{K}$, cf. Hestenes et al. (1952), known for its excellent regularization properties; see also Blanchard & Krämer (2016); Engl et al. (1996); Hanke (1995); Nemirovski (1986).[2]

The estimator is uniquely defined for every $m \leq n_*$, where n_* is the number of distinct non-zero eigenvalues of $\hat{K}$; see Proposition 1 in the Appendix B. It is also easy to see that for every $m \geq 1$,[3] we have $\hat{\beta}_m = \hat{P}_m(\hat{K})\hat{r}$ for a polynomial $\hat{P}_m(\hat{K}) = \sum_{j=1}^m a_j \hat{K}^{j-1}$ with coefficients $\mathbf{a} := (a_1, \dots, a_m)^\top$ solving the system of m linear equations

$$\mathbf{Ka} = \mathbf{r},$$

where $\mathbf{K} := \langle \hat{K}^j \hat{r}, \hat{K}^k \hat{r} \rangle_{1 \leq j,k \leq m}$ and $\mathbf{r} := \langle \hat{K}^j \hat{r}, \hat{r} \rangle_{1 \leq j \leq m}$. From the practical point of view, it is more efficient to use an iterative conjugate gradient algorithm that bypasses the (potentially unstable) matrix inversion with an iterative multiplication by the operator $\hat{K}$; see Algorithm 1 in section 3.

Notation For two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n \geq 1}$, we will use $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \leq c b_n$ for all $n \geq 1$. We will also use $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two real numbers, we use $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

2 Theoretical Properties

In this section, we will show that the functional PLS estimator achieves the (nearly) optimal convergence rate on a class of ellipsoids. We consider an early stopping rule for the estimator and show that it adapts to the complexity of the ellipsoid. Lastly, we study how rapidly, the number of selected components increases with the sample size and make some comparisons to the PCA estimator.

[2] The method of conjugate gradients is an efficient algorithm for solving high-dimensional systems of linear equations; see also (Nocedal & Wright, 1999, Chapter 5) and references therein.

[3] We also define $\hat{P}_0 = 0$ and $\hat{\beta}_0 = 0$.

2.1 Optimal Convergence Rates

Since the operators $K : \mathbb{H} \to \mathbb{H}$ and $\hat{K} : \mathbb{H} \to \mathbb{H}$ are self-adjoint and compact, by the spectral theorem

$$K = \sum_{j=1}^{\infty} \lambda_j v_j \otimes v_j \qquad \text{and} \qquad \hat{K} = \sum_{j=1}^{n} \hat{\lambda}_j \hat{v}_j \otimes \hat{v}_j,$$

where $\lambda_1 \geq \lambda_2 \geq \dots > 0$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_n \geq 0$ are the eigenvalues of K and $\hat{K}$ and $(v_j)_{j=1}^{\infty}$ and $(\hat{v}_j)_{j=1}^{n}$ are the corresponding eigenvectors; see Kress (1999), Theorem 15.16. Note that the sample covariance operator $\hat{K}$ is a finite-rank operator with at most $n_* \leq n$ distinct non-zero eigenvalues.

For a bounded and measurable function $\phi : \mathbb{R}_+ \to \mathbb{R}$, we define functions of operators through their spectral decompositions:

$$\phi(K) := \sum_{j=1}^{\infty} \phi(\lambda_j) v_j \otimes v_j \qquad \text{and} \qquad \phi(\hat{K}) := \sum_{j=1}^{n_*} \phi(\hat{\lambda}_j) \hat{v}_j \otimes \hat{v}_j.$$

The following inequalities for the operator norm will be often used:

$$\|\phi(K)\|_{\mathrm{op}} \leq \sup_{\lambda \in [0,\lambda_1]} |\phi(\lambda)| \qquad \text{and} \qquad \|\phi(\hat{K})\|_{\mathrm{op}} \leq \sup_{\lambda \in [0,\hat{\lambda}_1]} |\phi(\lambda)|, \tag{4}$$

where $\|A\|_{\mathrm{op}} = \sup_{\|x\|=1} \|Ax\|$.

We shall introduce several relatively mild assumptions on the distribution of the data next.

Assumption 1. *$(X_i, Y_i)_{i=1}^{n}$ are i.i.d. copies of (X, Y) with $\mathbf{E}[X] = 0$, $\mathbf{E}\|X\|^4 < \infty$, and $\mathbf{E}[\varepsilon^2|X] \leq \sigma^2 < \infty$.*

Assumption 1 imposes mild restrictions on the data-generating process. Note that $\mathbf{E}\|X\|^4 < \infty$ is satisfied when X is a Gaussian process in $\mathbb{H}$. It implies that K is a nuclear operator and, hence, compact.

Assumption 2. *The operator $K : \mathbb{H} \to \mathbb{H}$ does not have zero eigenvalues.*

Assumption 2 ensures that the slope parameter β is identified. It can be relaxed, in which case we could focus on the identified part of β in the orthogonal complement to the null space of K; see Babii & Florens (2017); Engl et al. (1996).

Assumption 3. *For some $\mu, R, C > 0$, the slope parameter β and the operator K belong to the class*

$$\mathcal{S}(\mu, R, C) = \left\{ \beta \in \mathbb{H},\ K : \mathbb{H} \to \mathbb{H} : \quad \sum_{j=1}^{\infty} \frac{\langle \beta, v_j \rangle^2}{\lambda_j^{2\mu}} \leq R^2, \quad \sum_{j=1}^{\infty} \lambda_j \leq C \right\}.$$

Assumption 3 describes the complexity of the ill-posed inverse problem in terms of the smoothness of β and the smoothing properties of the operator K. The parameter μ is known as the degree of ill-posedness. It restricts the rate of decline of the generalized Fourier coefficients $\langle \beta, v_j \rangle_{j \geq 1}$ relatively to the eigenvalues of K; see also Carrasco et al. (2007). Recall also that the summability of eigenvalues holds whenever $\mathbf{E}\|X\|^2 < \infty$.

Consider now the so-called residual polynomial $\hat{Q}_m(\lambda) = 1 - \lambda \hat{P}_m(\lambda)$, deriving its name from the identity $\hat{r} - \hat{K}\hat{\beta}_m = \hat{Q}_m(\hat{K})\hat{r}$. It is known that the polynomial, $\hat{Q}_m$, has m distinct real roots, denoted $\hat{\theta}_1 > \hat{\theta}_2 > \cdots > \hat{\theta}_m > 0$. The sum of inverse of these roots,

$$|\hat{Q}_m'(0)| = \sum_{j=1}^{m} \frac{1}{\hat{\theta}_j},$$

plays an important role in the analysis of the conjugate gradient regularization; see Lemma 6 in the Appendix B.

Our first result characterizes the convergence rate of the estimation and prediction errors of the PLS estimator.

Theorem 1. *Suppose that Assumptions 1, 2, and 3 are satisfied. Then for every $s \in [0, 1]$, we have*

$$\left\| K^s(\hat{\beta}_m - \beta) \right\|^2 = O_P\left(|\hat{Q}_m'(0)|^{2(1-s)} n^{-1} + |\hat{Q}_m'(0)|^{-2(\mu+s)} + |\hat{Q}_m'(0)|^{-2s} n^{-\mu \wedge 1} \right),$$

provided that $|\hat{Q}_m'(0)| = O_P(n^{1/2})$.

The proof of this and all the subsequent results appear in the Appendix A. Note that the last condition in Theorem 1 assumes that the number of components m does not increase too fast with the sample size and is not binding. In fact, it is optimal to have $|\hat{Q}_m'(0)| \sim n^{\frac{1}{2(\mu+1)}}$, in which case we obtain the following convergence rate

$$\left\|K^s(\hat{\beta}_m - \beta)\right\|^2 = O_P\left(n^{-\frac{\mu+s}{\mu+1}}\right).$$

When $s = 0$, this shows that the convergence rate of PLS in the Hilbert space norm is of order $n^{-\frac{\mu}{\mu+1}}$. On the other hand, when $s = 1/2$, we obtain the convergence rate of the out-of-sample prediction error, since

$$\mathbf{E}_X\langle X, \hat{\beta}_m - \beta\rangle^2 = \left\|K^{1/2}(\hat{\beta}_m - \beta)\right\|^2,$$

where $\mathbf{E}_X$ is taken with respect to X, independent of $(Y_i, X_i)_{i=1}^n$.

The following result shows that no estimator can achieve a faster than $n^{-\frac{\mu+s}{\mu+1}}\log^{-b} n$ rate on the class $\mathcal{S}(\mu, R, C)$.

Theorem 2. *For every* $s \in [0, 1/2]$, *there exists* $A < \infty$ *such that*

$$\liminf_{n\to\infty} \inf_{\hat{\beta}} \sup_{(\beta,K)\in\mathcal{S}(\mu,R,C)} \Pr\left(\left\|K^s(\hat{\beta} - \beta)\right\| \geq An^{-\frac{\mu+s}{2(\mu+1)}}\log^{-b/2} n\right) > 0,$$

where $b > 2(\mu + s)$ *and the infimum is over all estimators.*

Therefore, we conclude that the PLS estimator $\hat{\beta}_m$ achieves the (nearly) optimal convergence rate on $\mathcal{S}(\mu, R, C)$, simultaneously for the estimation and prediction errors.[4]

2.2 Adaptive PLS estimator

Next, we look at the adaptive PLS estimator, where the number of components is selected using the data-driven rule described in the following assumption.

[4] It is possible to avoid the $1/\log n$ factor by considering the larger class of Hilbert–Schmidt operators.

Assumption 4. *We select $\hat{m}$ such that*

$$\min\left\{m \geq 0 : \left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \leq \tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}}\right\}.$$

for a sufficiently large $\tau > 1$ and some $\delta \in (0,1)$.

Assumption 4 states that the PLS iterations stop at the first value $\hat{m}$ for which the norm of residual is smaller than a certain threshold. Note that the number of iterations is finite since $\hat{m} \leq n_*$, where n_* is the number of distinct non-zero eigenvalues of $\hat{K}$; see Proposition 1 in the Appendix B. In fact, the residual is zero for all $m \geq n_*$.

The following result shows that the data-driven rule in Assumption 4 is adaptive to the unknown degree of ill-posedness $\mu > 0$.[5]

Theorem 3. *Suppose that Assumptions 1, 2, 3, and 4 hold with $\delta \geq 1/n$. Then*

$$\left\|K^s(\hat{\beta}_{\hat{m}} - \beta)\right\|^2 = O\left((\delta n)^{-\frac{\mu+s}{\mu+1}}\right)$$

with probability at least $1-\delta$ for every $s \in [0,1]$.

Taking $\delta_n = 1/\log n$ in Assumption 4, we obtain from Theorem 3 the convergence rate of the estimation and prediction errors of PLS with the early stopping rule:

$$\left\|K^s(\hat{\beta}_{\hat{m}} - \beta)\right\|^2 = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{\mu+s}{\mu+1}}\right).$$

Therefore, the adaptive PLS achieves the (nearly) optimal convergence rate simultaneously for the estimation and prediction errors without knowing the degree of ill-posedness $\mu > 0$.

2.3 Number of Selected Components

In this section, we look at how rapidly the number of components selected by the early stopping rule in the Assumption 4 increases with the sample size. First, we consider a

[5] See also Blanchard et al. (2018); Blanchard & Krämer (2016); Blanchard & Mathé (2012) for the analysis of early stopping rules in various ill-posed inverse problems and Cai & Yuan (2012); Comte & Johannes (2012) for the adaptive functional linear regression model.

somewhat conservative bound that does not impose any assumptions on the spectrum of the operator K.

Theorem 4. *Suppose that Assumptions 1, 2, 3, and 4 are satisfied with $\delta \geq 1/n$ and $\mu \geq 1$. Then with probability at least $1-\delta$*

$$\hat{m} = O\left((n\delta)^{\frac{1}{4(\mu+1)}}\right).$$

Taking $\delta = 1/\log n$, we obtain from Theorem 4 that $\hat{m} = O_P\left((n/\log n)^{\frac{1}{4(\mu+1)}}\right)$. Next, we consider sharper estimates under additional assumptions imposed on the spectrum of the operator K.

Theorem 5. *Suppose that Assumptions 1, 2, 3, and 4 are satisfied with $\delta \geq e/n$ and $\mu \geq 1$. Then with probability at least $1-\delta$*

(i) If $\lambda_j = O(j^{-2\kappa})$ for some $\kappa > 0$, then

$$\hat{m} = O\left((n\delta)^{\frac{1}{4(\kappa+1)(\mu+1)}}\right).$$

(ii) If $\lambda_j = O(q^j)$ for some $q \in (0,1)$, then

$$\hat{m} = O\left(\log(n\delta)\right).$$

Theorem 5 shows that if the eigenvalues decline polynomially fast, then the selected number of components is $\hat{m} = O_P(n/\log n)^{\frac{1}{4(\kappa+1)(\mu+1)}}$ while in the case of the geometric decline, the number of selected components increases slowly with the sample size. Therefore, the adaptive stopping rule will select a smaller number of components if the eigenvalues of the operator K decline faster and vice versa.

2.4 Comparison to PCA

It appears challenging to compare directly the risks of functional PCA and PLS. In this section, we shed some light on the behavior of functional PLS relative to PCA. We will show that for the same fixed number of components m, PLS fits the empirical moment better than PCA, hence, it may require a smaller number of components to obtain a comparable fit. We also show that the regularization bias part of the estimation and prediction risk of PLS is smaller than the one of the PCA. Therefore, the adaptive PLS basis is better suited for approximating the slope coefficient.

In what follows, we will use

$$\hat{\beta}_m^{\mathrm{PLS}} = \sum_{j=1}^{n_*} \hat{P}_m(\hat{\lambda}_j)\langle \hat{r}, \hat{v}_j\rangle \hat{v}_j \qquad \text{and} \qquad \hat{\beta}_m^{\mathrm{PCA}} = \sum_{j=1}^{m} \frac{1}{\hat{\lambda}_j}\langle \hat{r}, \hat{v}_j\rangle \hat{v}_j$$

to denote the functional PLS and PCA estimators. Note that the PLS estimator uses supervised regularization $\hat{P}_m$ while for the PCA estimator the regularization is fixed to select the terms related to the inverse of the largest m eigenvalues of $\hat{K}$. We will also use

$$\beta_m^{\mathrm{PLS}} = \sum_{j=1}^{\infty} P_m(\lambda_j)\langle r, v_j\rangle v_j \qquad \text{and} \qquad \beta_m^{\mathrm{PCA}} = \sum_{j=1}^{m} \lambda_j^{-1}\langle r, v_j\rangle v_j$$

to denote the population counterparts.

Theorem 6. *For every $m \leq n_*$,*

$$\left\| \hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PLS}} \right\| \leq \left\| \hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PCA}} \right\|.$$

and

$$\left\| K^s(\beta_m^{\mathrm{PLS}} - \beta) \right\| \leq \left\| K^s(\beta_m^{\mathrm{PCA}} - \beta) \right\|, \qquad \forall s \in [0,1].$$

The first part of Theorem 6 shows that the PLS estimator fits the data better than PCA for the same number of components $1 \leq m \leq n_*$. This is a functional version of a result discussed in Blazere et al. (2014). For the second part of Theorem 6, it is worth recalling that the estimation and prediction errors in Theorem 1 can be decomposed as

$$K^s\left(\hat{\beta}_m^{\mathrm{PLS}} - \beta\right) = K^s\left(\hat{\beta}_m^{\mathrm{PLS}} - \beta_m^{\mathrm{PLS}}\right) + K^s\left(\beta_m^{\mathrm{PLS}} - \beta\right), \qquad s \in \{0, 1/2\},$$

where the second term is the so-called regularization bias. This shows that the PLS basis is better suited for approximating the slope β than the PCA basis.

3 Monte Carlo Experiments

In this section, we set up several Monte Carlo experiments to evaluate the finite sample performance of the PLS estimator.

We simulate the i.i.d. samples $(Y_i, X_i)_{i=1}^n$ as follows

$$Y_i = \int_0^1 X_i(s)\beta(s)ds + \varepsilon_i, \qquad \varepsilon_i \sim_{\text{i.i.d.}} N(0,1),$$

where the predictors belong to the Hilbert space of square-integrable functions with respect to the Lebesgue measure, denoted $\mathbb{H} = L^2[0,1]$. The functional predictor is generated as

$$X_i(s) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} u_j v_j(s), \qquad u_j \sim_{\text{i.i.d.}} N(0,1).$$

The slope parameter $\beta \in L^2[0,1]$ and the spectrum of the operator $(\lambda_j, v_j)_{j\geq 1}$ correspond to one of the following four models:

Model 1: $\beta(s) = \sum_{j=1}^{\infty} \beta_j v_j(s)$ with $\beta_j = 4(-1)^{j+1} j^{-2}$, $v_j(s) = \sqrt{2}\cos((j-1)\pi s)$, and $\lambda_j = j^{-2}, \forall j \geq 1$; see Hall & Horowitz (2007).

Model 2: same as Model 1, but with $\beta_5 = 4$.

Model 3: same as Model 1, but with $\beta_5 = \beta_{10} = 4$

Model 4: same as Model 1, but with $\lambda_j = 0.5^j$.

Note that the first few high-variance principal components terms are sufficient to capture most of nonlinearities in Model 1. Therefore, this design favors strongly PCA. On the other hand, the low-variance components are important in the slope parameter β for Model 2 and even more so for Model 3; cf. Jolliffe (1982). Model 4 is an example of severely ill-posed problem with eigenvalues declining rapidly at the geometric rate. We compute the PLS estimator using the Algorithm 1 which is a very efficient way to compute the solution of equation (3); see Nocedal & Wright (1999), Algorithm 5.2. It bypasses the operator inversion with an iterative multiplication by $\hat{K}$ and can be understood as a variation of the conjugate gradient algorithm popular for solving high-dimensional systems of linear equations with a symmetric matrix.

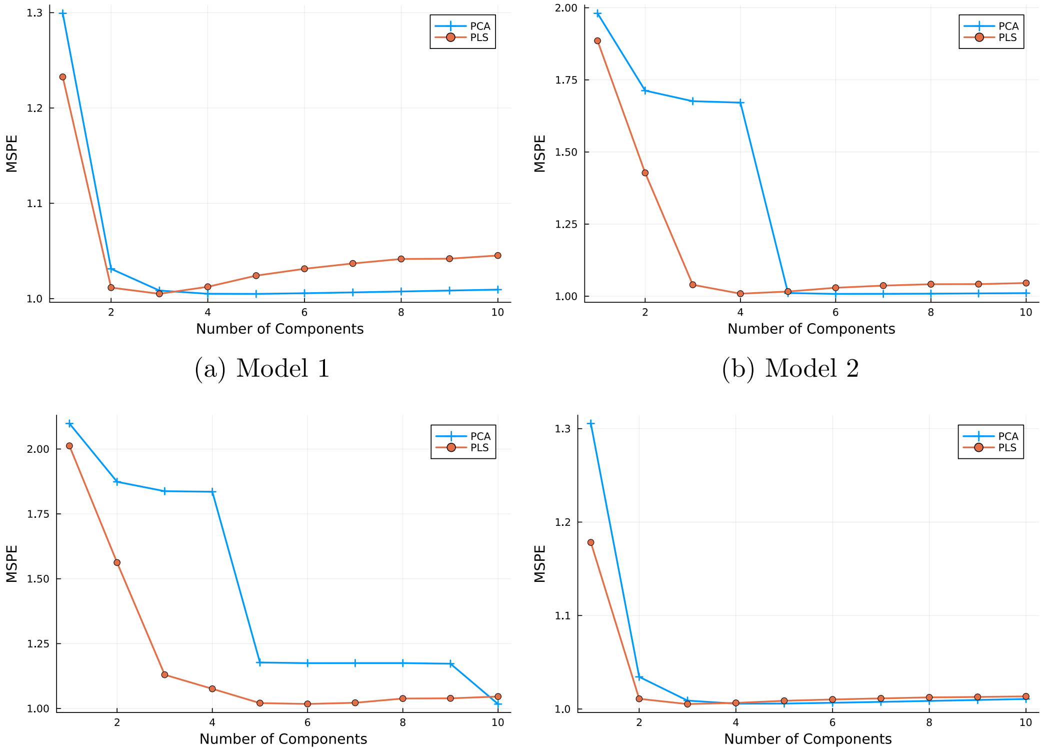
Algorithm 1: PLS algorithm for solving $\hat{K}\hat{\beta} = \hat{r}$.

```
Result: $\hat{\beta}_m$
Initialisation: $\hat{\beta}_0 = 0$, $e_0 = \hat{K}\hat{\beta}_0 - \hat{r}$, $d_0 = -e_0$;
for $j = 0, 1, \dots, m-1$ do
    1. Compute the step size: $\alpha_j = \frac{\langle e_j, e_j \rangle}{\langle d_j, \hat{K} d_j \rangle}$;
    2. Update the slope coefficient: $\hat{\beta}_{j+1} = \hat{\beta}_j + \alpha_j d_j$;
    3. Update the residual: $e_{j+1} = e_j + \alpha_j \hat{K} d_j$;
    4. Compute the step size for the conjugate direction update: $\gamma_{j+1} = \frac{\langle e_{j+1}, e_{j+1} \rangle}{\langle e_j, e_j \rangle}$;
    5. Update the conjugate direction vector: $d_{j+1} = -e_{j+1} + \gamma_{j+1} d_j$;
end
```

The integrals in inner products and the operator K are discretized using a simple approximation with Riemann sum on a grid of $T = 100$ equidistant points in $[0, 1]$ and the infinite sums are truncated at $J = 50$. The experiments feature $5,000$ replications, where samples of size $n = 1,000$ are generated in each replication. For each experiment, the mean-squared prediction error (MSPE) is computed as

$$\text{MSPE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, \hat{\beta} \rangle)^2,$$

Figure 1: Mean-Squared Prediction Error (MSPE) of PLS (orange circles) and PCA (blue crosses) using the first m components, calculated from $5,000$ samples of size $n = 1,000$.



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

where $\hat{\beta}$ is obtained from an auxiliary sample of size n, independent of $(Y_i, X_i)_{i=1}^{n}$.

Figure 1 displays the MSPE using the first m components using PLS (orange) or PCA (blue); see also Table 1 for the exact values of the MSPE. The PLS estimator achieves the lowest value of MSPE across all designs and it usually does so with a smaller number of components. Remarkably, even for Model 1, where PCA is expected to perform extremely well, the PLS estimator is able to reduce the MSPE faster for the first few components. The early stopping of PLS in this case is important since it starts overfitting for larger number of components. For Model 2 and Model 3, where the low-variance components are important, the PLS estimator outperforms PCA by a large margin. Lastly, PLS outperforms PCA in the severely ill-posed case (Model 4) and has a comparable to PCA overfitting pattern.

Table 1: Mean-Squared Prediction Error (MSPE) of PLS and PCA using the first m components, calculated from $1,000$ samples of size $n = 1,000$.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	1.299	1.233	1.981	1.886	2.098	2.012	1.305	1.178
2	1.031	1.012	1.713	1.428	1.874	1.562	1.034	1.011
3	1.008	1.005	1.676	1.040	1.838	1.130	1.009	1.005
4	1.005	1.012	1.671	1.009	1.835	1.076	1.006	1.007
5	1.005	1.024	1.011	1.016	1.177	1.021	1.006	1.009
6	1.006	1.031	1.008	1.029	1.175	1.018	1.007	1.010
7	1.007	1.037	1.008	1.037	1.175	1.022	1.008	1.011
8	1.007	1.042	1.009	1.042	1.175	1.038	1.009	1.012
9	1.008	1.042	1.010	1.042	1.173	1.039	1.010	1.013
10	1.009	1.045	1.011	1.046	1.017	1.046	1.011	1.013

Next, we look at the mean-squared error (MSE) for the slope parameter β. Table 2 shows that the MSE displays the same pattern declining faster with m for the PLS estimator. The optimal number of the PLS vs. PCA components are respectively: 3 vs. 5 (Model 1), 4 vs. 6 (Models 2), 5 vs. 10 (Model 3), and 3 vs. 4 (Model 4). The PLS also achieves lower MSE for Model 4 with just 3 components. Note that for Model 3, the number of components needed for PCA is two times larger.

Tables 3 and 4 display the bias and the variance MSE components. We can see that the remarkable MSE performance of PLS with a smaller number of components comes from its smaller bias and comparable variance when m is small, confirming the result of Theorem 6. The simulation results also underscore the importance of early stopping for the PLS since the method overfits if the number of components gets larger. To conclude, the results of the experiments confirm our theoretical results and illustrate that the supervised PLS bases are better suited for representing the slope parameter β and for predicting the response variable Y.

Table 2: Mean-Squared Error (MSE) of PLS and PCA using the first m components, calculated from $5,000$ samples of size $n = 1,000$.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	1.394	1.116	17.780	16.835	33.851	32.787	1.406	0.874
2	0.346	0.150	16.725	9.555	32.966	23.37	0.341	0.132
3	0.144	0.104	16.387	0.602	32.625	9.249	0.144	0.077
4	0.092	0.626	16.25	0.308	32.528	6.460	0.092	0.144
5	0.089	2.638	0.309	1.448	16.573	1.806	0.095	0.426
6	0.111	6.236	0.209	4.566	16.469	2.155	0.147	1.035
7	0.152	11.207	0.221	9.740	16.449	5.027	0.266	2.233
8	0.211	16.994	0.261	16.08	16.391	13.631	0.515	4.577
9	0.287	17.459	0.330	16.527	16.057	14.176	1.027	6.222
10	0.381	23.358	0.416	22.819	1.350	20.182	2.043	9.239

Table 3: Squared Bias part of MSE of PLS and PCA using the first m components, calculated from $5,000$ samples of size $n = 1,000$.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	1.381	1.104	17.767	16.823	33.837	32.775	1.379	0.856
2	0.331	0.132	16.707	9.509	32.947	23.311	0.323	0.117
3	0.123	0.028	16.347	0.324	32.583	8.914	0.125	0.041
4	0.058	0.022	16.113	0.006	32.387	6.046	0.061	0.020
5	0.031	0.017	0.021	0.008	16.276	0.063	0.034	0.013
6	0.019	0.013	0.009	0.004	16.244	0.015	0.021	0.009
7	0.012	0.008	0.008	0.004	16.182	0.018	0.014	0.007
8	0.008	0.006	0.004	0.005	15.999	0.011	0.010	0.006
9	0.006	0.005	0.004	0.005	15.209	0.010	0.007	0.006
10	0.004	0.005	0.002	0.006	0.118	0.009	0.005	0.005

Table 4: Variance part of MSE of PLS and PCA using the first m components, calculated from $5,000$ samples of size $n = 1,000$.

	Model 1		Model 2		Model 3		Model 4	
m	PCA	PLS	PCA	PLS	PCA	PLS	PCA	PLS
1	0.013	0.011	0.013	0.012	0.013	0.012	0.027	0.019
2	0.014	0.018	0.019	0.046	0.019	0.059	0.018	0.016
3	0.021	0.076	0.040	0.278	0.042	0.335	0.019	0.037
4	0.034	0.604	0.137	0.302	0.140	0.415	0.031	0.124
5	0.058	2.621	0.288	1.440	0.297	1.743	0.061	0.413
6	0.093	6.223	0.200	4.562	0.225	2.140	0.125	1.026
7	0.140	11.199	0.213	9.736	0.267	5.009	0.252	2.226
8	0.203	16.988	0.258	16.075	0.392	13.62	0.505	4.571
9	0.282	17.453	0.326	16.523	0.848	14.165	1.020	6.216
10	0.377	23.353	0.414	22.813	1.232	20.172	2.037	9.234

A Proofs

In this appendix, we consider several auxiliary lemmas and provide detailed proofs for the main results. In what follows, for $k \in \mathbb{Z}$, consider the following measure

$$\hat{\mu}_k = \sum_{j=1}^{n_*} \hat{\lambda}_j^k \langle \hat{r}, \hat{v}_j \rangle^2 \delta_{\hat{\lambda}_j},$$

where δ_x is the Dirac measure at $x \in \mathbb{R}$. For $\phi, \psi : [0, \hat{\lambda}_1] \to \mathbb{R}$, define

$$\begin{aligned} [\phi, \psi]_k &= \int_0^\infty \phi(\lambda)\psi(\lambda) \mathrm{d}\hat{\mu}_k(\lambda) \\ &= \sum_{j=1}^{n_*} \phi(\hat{\lambda}_j)\psi(\hat{\lambda}_j)\hat{\lambda}_j^k \langle \hat{r}, \hat{v}_j \rangle^2. \end{aligned} \tag{5}$$

Lastly, let

$$\Pi_a = \sum_{j:\hat{\lambda}_j \leq a} \hat{v}_j \otimes \hat{v}_j$$

be the orthogonal projection operators on the eigenspaces of $\hat{K}$ corresponding to eigenvalues smaller or equal to a.

The following lemma allows us to control the residuals of the PLS estimator.

Lemma 1. *Suppose that Assumptions 1, 2, and 3 are satisfied Then for every* $1 \leq m \leq n_*$ *and* $\gamma \in [1/n, 1)$

$$\left\| \hat{r} - \hat{K}\hat{\beta}_m \right\| \lesssim \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} + |\hat{Q}_m'(0)|^{-1} \sqrt{\frac{2\mathbf{E}\|X\|^4}{\gamma n}} + |\hat{Q}_m'(0)|^{-(\mu+1)}$$

on an event with probability at least $1 - \gamma$.

Proof of Lemma 1. Let $\varphi_m(\lambda) := \hat{Q}_m(\lambda)\sqrt{\hat{\theta}_m/(\hat{\theta}_m - \lambda)}$. We will first show that the following inequality holds

$$\begin{aligned} \left\| \hat{r} - \hat{K}\hat{\beta}_m \right\| &= \left\| \hat{Q}_m(\hat{K})\hat{r} \right\| \\ &\leq \left\| \Pi_{\hat{\theta}_m} \varphi_m(\hat{K})\hat{r} \right\|, \end{aligned}$$

where the first line uses $\hat{\beta}_m = \hat{P}_m(\hat{K})\hat{r}$ and $\hat{Q}_m(\lambda) = 1 - \lambda\hat{P}_m(\lambda)$. The inequality can be deduced from the proof of Theorem 7.9 in Engl et al. (1996). For completeness, we provide an argument suitably tailored to our setting below. By Lemma 6 (iii) and (v) since the polynomials $(\hat{Q}_m)_{m\geq 0}$ are orthogonal with respect to $[.,.]_1$, see equation (5), we get for $m \geq 1$

$$\begin{aligned} 0 &= \int_0^\infty \hat{Q}_m(\lambda)\hat{Q}_{m-1}(\lambda)\mathrm{d}\hat{\mu}_1(\lambda) \\ &= \hat{\theta}_m \int_0^\infty \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) \\ &= \hat{\theta}_m \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) + \hat{\theta}_m \int_{\hat{\theta}_m}^\infty \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda). \end{aligned}$$

Since $\hat{\theta}_m > 0$, by Lemma 6 (i), this shows that

$$\int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m - \lambda}\mathrm{d}\hat{\mu}_0(\lambda) = \int_{\hat{\theta}_m}^\infty \hat{Q}_m^2(\lambda)\frac{\lambda}{\lambda - \hat{\theta}_m}\mathrm{d}\hat{\mu}_0(\lambda)$$

and so by equation (15)

$$\begin{aligned}
\left\|\hat{Q}_m(\hat{K})\hat{r}\right\|^2 &= \int_0^\infty \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) \\
&= \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) + \int_{\hat{\theta}_m}^\infty \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) \\
&\leq \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) + \int_{\hat{\theta}_m}^\infty \hat{Q}_m^2(\lambda)\frac{\lambda}{\lambda-\hat{\theta}_m}\mathrm{d}\hat{\mu}_0(\lambda) \\
&= \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\mathrm{d}\hat{\mu}_0(\lambda) + \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\lambda}{\hat{\theta}_m-\lambda}\mathrm{d}\hat{\mu}_0(\lambda) \\
&= \int_0^{\hat{\theta}_m} \hat{Q}_m^2(\lambda)\frac{\hat{\theta}_m}{\hat{\theta}_m-\lambda}\mathrm{d}\hat{\mu}_0(\lambda) \\
&= \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{r}\right\|^2
\end{aligned}$$

where the third line follows since $1 \leq \lambda/(\lambda-\hat{\theta}_m)$ for all $\lambda \geq \hat{\theta}_m$.

Therefore,

$$\begin{aligned}
\left\|\hat{r}-\hat{K}\hat{\beta}_m\right\| &\leq \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{r}\right\| \\
&\leq \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\beta\right\| + \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})(\hat{r}-\hat{K}\beta)\right\|.
\end{aligned}$$

Under Assumption 3, $\beta = K^\mu w$ with $\|w\| \leq R$, so that

$$\begin{aligned}
\left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\beta\right\| &= \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}K^\mu w\right\| \\
&= \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\hat{K}^\mu w\right\| + \left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})\hat{K}\left[K^\mu - \hat{K}^\mu\right]w\right\| \\
&\leq \sup_{\lambda\in[0,\hat{\theta}_m]}\left|\varphi_m(\lambda)\lambda^{1+\mu}\right| R + \sup_{\lambda\in[0,\hat{\theta}_m]}|\varphi_m(\lambda)\lambda| \left\|\hat{K}^\mu - K^\mu\right\|_{\mathrm{op}} R \\
&\leq (2\mu+2)^{\mu+1}|\hat{Q}_m'(0)|^{-(\mu+1)}R \\
&\qquad + 2|\hat{Q}_m'(0)|^{-1}(c_\mu\mathbf{1}_{\mu\leq 1} + \mu\nu^{\mu-1}\mathbf{1}_{\mu>1})\left(\frac{2\mathbf{E}\|X\|^4}{\gamma n}\right)^{\frac{\mu\wedge 1}{2}} R,
\end{aligned}$$

where the last line follows by Lemma 6 (vi) and the inequality in equation (16) on an event with probability at least $1-\gamma$. By Lemma 4

$$\begin{aligned}\nu = \|\hat{K}\|_{\mathrm{op}} \vee \|K\|_{\mathrm{op}} &\leq \|K\|_{\mathrm{op}} + \left\|\hat{K}-K\right\|_{\mathrm{op}} \\ &\leq \lambda_1 + \sqrt{\frac{2\mathbf{E}\|X\|^4}{\gamma n}} \lesssim 1\end{aligned} \tag{6}$$

on an event with probability at least $1-\gamma$.

Lastly,

$$\begin{aligned}\left\|\Pi_{\hat{\theta}_m}\varphi_m(\hat{K})(\hat{r}-\hat{K}\beta)\right\| &\leq \sup_{\lambda\in[0,\hat{\theta}_m]}|\varphi_m(\lambda)|\left\|\hat{r}-\hat{K}\beta\right\| \\ &= \sup_{\lambda\in[0,\hat{\theta}_m]}\left|\hat{Q}_m(\lambda)\sqrt{\hat{\theta}_m/(\hat{\theta}_m-\lambda)}\right|\left\|\hat{r}-\hat{K}\beta\right\| \\ &\leq \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}},\end{aligned}$$

where the last line follows by Lemma 6 (vi) with $\delta=0$ and Lemma 4. □

Proof of Theorem 1. Take any $m \leq n_*$, $\gamma\in(0,1)$, and let $a>0$ be such that $a \leq |\hat{Q}_m'(0)|^{-1}$. By Lemma 6 (iv) this ensures that $a\leq\hat{\theta}_m$ which we will use repeatedly in the proof.

Decompose

$$\hat{\beta}_m-\beta = \Pi_a\hat{P}_m(\hat{K})(\hat{r}-\hat{K}\beta) + \Pi_a\left[\hat{P}_m(\hat{K})\hat{K}-I\right]\beta + \Pi_a^\perp(\hat{\beta}_m-\beta),$$

where $\Pi_a=\sum_{j:\hat{\lambda}_j\leq a}\hat{v}_j\otimes\hat{v}_j$ and $\Pi_a^\perp = I-\Pi_a$. Then for $s\in[0,1]$, we have

$$\begin{aligned}\left\|\hat{K}^s(\hat{\beta}_m-\beta)\right\| &\leq \left\|\Pi_a\hat{K}^s\hat{P}_m(\hat{K})(\hat{r}-\hat{K}\beta)\right\| + \left\|\Pi_a\hat{K}^s\hat{Q}_m(\hat{K})\beta\right\| + \left\|\Pi_a^\perp\hat{K}^s(\hat{\beta}_m-\beta)\right\| \\ &=: I+II+III.\end{aligned}$$

We will derive an upper bound for each of these three terms separately. For the first term, note that for every $s \in [0,1]$,

$$\begin{aligned} I &\leq \left\| \Pi_a \hat{K}^s \hat{P}_m(\hat{K}) \right\| \left\| \hat{r} - \hat{K}\beta \right\| \\ &\leq \sup_{\lambda \in [0,a]} |\lambda^s \hat{P}_m(\lambda)| \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} \\ &\leq a^s |\hat{Q}_m'(0)| \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}}, \end{aligned}$$

where the second line follows on an event with probability at least $1-\gamma$ by Lemma 4 and equation (4), and the last one by the convexity of $\hat{Q}_m$ on $[0,a]$:

$$\hat{P}_m(\lambda) = \frac{1 - \hat{Q}_m(\lambda)}{\lambda} \leq -\hat{Q}_m'(0),$$

and $\hat{Q}_m(\lambda) \leq \hat{Q}_m(0) = 1$ for every $\lambda \in [0,a]$; see Lemma 6 (ii).

For the second term, under Assumption 3, we have $\beta = K^\mu w$ with $\|w\| \leq R$, so that

$$\begin{aligned} II &= \left\| \Pi_a \hat{K}^s \hat{Q}_m(\hat{K}) K^\mu w \right\| \\ &\leq \left\| \Pi_a \hat{K}^s \hat{Q}_m(\hat{K}) \hat{K}^\mu w \right\| + \left\| \Pi_a \hat{K}^s \hat{Q}_m(\hat{K}) \left[K^\mu - \hat{K}^\mu \right] w \right\| \\ &\leq \sup_{\lambda \in [0,a]} |\lambda^{\mu+s} \hat{Q}_m(\lambda)| R + \sup_{\lambda \in [0,a]} |\lambda^s \hat{Q}_m(\lambda)| \left\| K^\mu - \hat{K}^\mu \right\|_{\mathrm{op}} R \\ &\leq a^{\mu+s} R + a^s (c_\mu \mathbf{1}_{\mu \leq 1} + \mu \nu^{\mu-1} \mathbf{1}_{\mu > 1}) \left(\frac{2\mathbf{E}\|X\|^4}{\gamma n} \right)^{\frac{\mu \wedge 1}{2}} R, \end{aligned}$$

where the last line follows since $|\hat{Q}_m(\lambda)| \leq 1$ by Lemma 6 (ii) and equation (16) on an event with probability at least $1-\gamma$.

Lastly, let $\hat{K}^+ = \sum_{j=1}^{n_*} \hat{v}_j \otimes \hat{v}_j / \hat{\lambda}_j$ be the generalized inverse of $\hat{K}$. Then we bound the second term as follows

$$\begin{aligned} III &\leq \left\|\Pi_a^{\perp}\hat{K}^s\hat{K}^+\right\| \left\|\hat{K}(\hat{\beta}_m - \beta)\right\| \\ &\leq \sup_{\lambda \geq a} \lambda^{s-1} \left\|(\hat{K}\hat{\beta}_m - \hat{r}) + (\hat{r} - \hat{K}\beta)\right\| \\ &\leq a^{s-1}\left\{\left\|\hat{K}\hat{\beta}_m - \hat{r}\right\| + \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}}\right\}, \end{aligned}$$

where we use Lemma 4 on an event with probability at least $1-\gamma$. Combining the three bounds, we obtain for $m \leq n_*$

$$\left\|\hat{K}^s(\hat{\beta}_m - \beta)\right\| \lesssim a^{s-1}\left\{\left\|\hat{K}\hat{\beta}_m - \hat{r}\right\| + (\gamma n)^{-1/2}\right\} + a^{\mu+s} + a^s(\gamma n)^{-\frac{\mu\wedge 1}{2}} + a^s|\hat{Q}_m'(0)|(\gamma n)^{-1/2}, \tag{7}$$

where we use $\nu \lesssim 1$; cf. equation (6). Taking $a = |\hat{Q}_m'(0)|^{-1}$, this gives

$$\left\|\hat{K}^s(\hat{\beta}_m - \beta)\right\| = O_P\left(|\hat{Q}_m'(0)|^{-s+1}\left\{\left\|\hat{K}\hat{\beta}_m - \hat{r}\right\| + n^{-1/2}\right\} + |\hat{Q}_m'(0)|^{-(\mu+s)} + |\hat{Q}_m'(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right).$$

By Lemma 1

$$\left\|\hat{K}\hat{\beta}_m - \hat{r}\right\| = O_P\left(n^{-1/2} + |\hat{Q}_m'(0)|^{-1}n^{-1/2} + |\hat{Q}_m'(0)|^{-(\mu+1)}\right).$$

Therefore,

$$\left\|\hat{K}^s(\hat{\beta}_m - \beta)\right\| = O_P\left(|\hat{Q}_m'(0)|^{-s+1}n^{-1/2} + |\hat{Q}_m'(0)|^{-(\mu+s)} + |\hat{Q}_m'(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right), \qquad \forall s \in [0,1].$$

This proves the result if $s = 0$. If $s \in (0,1]$, then

$$\begin{aligned}\left\|K^s(\hat{\beta}_m-\beta)\right\| &= \left\|\hat{K}^s(\hat{\beta}_m-\beta)+(K^s-\hat{K}^s)(\hat{\beta}_m-\beta)\right\| \\ &\leq \left\|\hat{K}^s(\hat{\beta}_m-\beta)\right\| + \left\|\hat{K}^s-K^s\right\|_{\mathrm{op}}\left\|\hat{\beta}_m-\beta\right\| \\ &= O_P\left(|\hat{Q}_m'(0)|^{-s+1}n^{-1/2}+|\hat{Q}_m'(0)|^{-(\mu+s)}+|\hat{Q}_m'(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right) \\ &\quad + O_P\left(|\hat{Q}_m'(0)|n^{-\frac{1+s}{2}}+|\hat{Q}_m'(0)|^{-\mu}n^{-\frac{s}{2}}+n^{-\frac{s+\mu\wedge 1}{2}}\right) \\ &= O_P\left(|\hat{Q}_m'(0)|^{-s+1}n^{-1/2}+|\hat{Q}_m'(0)|^{-(\mu+s)}+|\hat{Q}_m'(0)|^{-s}n^{-\frac{\mu\wedge 1}{2}}\right),\end{aligned}$$

provided that $|\hat{Q}_m'(0)|=O_P(n^{1/2})$. □

Proof of Theorem 2. We adopt an approach similar to Cai & Yuan (2012) Theorem 1; see also Tsybakov (2009), Chapter 2. Recall that the lower bound for a restricted class of models yields the lower bound for the general case. Therefore, we can assume without loss of generality that $\varepsilon_i|X_i \sim N(0,\sigma^2)$ and $K:\mathbb{H}\to\mathbb{H}$ has a spectral decomposition $(\lambda_j,v_j)_{j\geq 1}$ with $\lambda_1=1$ and $\lambda_j=1/(j\log^a j)$ for $j=2,3,\dots$ for some $a>1$. We will also consider the family of slope parameters

$$\beta_\theta = Rm^{-1/2}\sum_{l=m+1}^{2m}\theta_l\lambda_l^\mu v_l, \qquad \theta=(\theta_{m+1},\dots,\theta_{2m})\in\{0,1\}^m$$

for some $R>0$ and m specified below. It is easy to see that by the orthonormality of $(v_l)_{l\geq 1}$

$$\sum_{j=1}^{\infty}\frac{\langle\beta_\theta,v_j\rangle^2}{\lambda_j^{2\mu}} = \frac{R^2}{m}\sum_{j=m+1}^{2m}\theta_j^2 \leq R^2$$

and that

$$\sum_{j=1}^{\infty}\lambda_j = 1+\sum_{j=2}^{\infty}\frac{1}{j\log^a j}\leq C$$

for some $C>0$. Therefore, $(\beta_\theta,K)\in\mathcal{S}(\mu,R,C),\forall\theta\in\{0,1\}^m$, cf. Assumption 3.

Let $H(\theta,\theta')=\sum_{j=1}^m \mathbf{1}\{\theta_j\neq\theta_j'\}$ be the Hamming distance between the binary sequences $\theta,\theta'\in\{0,1\}^m$. By the Varshamov-Gilbert bound, see Tsybakov (2009), Lemma 2.9, if

$m \geq 8$, there exists $\{\theta^{(0)}, \ldots, \theta^{(M)}\} \subset \{0,1\}^m$ such that

(a) $\theta^{(0)} = (0, \ldots, 0)$;

(b) $H(\theta^{(j)}, \theta^{(k)}) \geq \frac{m}{8}, \forall\, 0 \leq j < k \leq M$;

(c) $M \geq 2^{m/8}$.

For every $A > 0$,

$$\begin{aligned}
&\sup_{(\beta,K)\in\mathcal{S}(\mu,R,C)} \Pr\left(\left\|K^s(\hat{\beta}-\beta)\right\| \geq An^{-\frac{\mu+s}{2(\mu+1)}}(\log n)^{-a(\mu+s)}\right) \\
&\geq \max_{\theta\in\{\theta^{(0)},\ldots,\theta^{(M)}\}} \Pr\left(\left\|K^s(\hat{\beta}-\beta_\theta)\right\| \geq An^{-\frac{\mu+s}{2(\mu+1)}}(\log n)^{-a(\mu+s)}\right).
\end{aligned} \tag{8}$$

To obtain the lower bound for the right-hand side of the equation (8), we will use Tsybakov (2009), Theorem 2.5 and a specific choice of $A < \infty$. To that end, we need to check the following conditions:

(i) $\|K^s(\beta_{\theta^{(j)}} - \beta_{\theta^{(k)}})\| \geq 2An^{-\frac{\mu+s}{2(\mu+1)}}(\log n)^{-a(\mu+s)}$ for all $0 \leq j < k \leq M$;

(ii) $P_j << P_0, \forall j = 1, \ldots, M$, where P_j denotes the distribution of $(Y_i, X_i)_{i\geq 1}$ for the slope parameter $\beta_{\theta^{(j)}}$;

(iii) For $\alpha \in (0, 1/8)$,

$$\frac{1}{M}\sum_{j=1}^{M} KL(P_j, P_0) \leq \alpha \log M,$$

where KL is the Kullback-Leibler divergence between P_j and P_0.

For the first condition, note that since $H(\theta^{(j)},\theta^{(k)})=\sum_{l=m+1}^{2m}(\theta_l^{(j)}-\theta_l^{(k)})^2$, we have

$$\begin{aligned}
\|K^s(\beta_{\theta^{(j)}}-\beta_{\theta^{(k)}})\|^2 &= \left\|Rm^{-1/2}\sum_{l=m+1}^{2m}(\theta_l^{(j)}-\theta_l^{(k)})\lambda_l^{\mu+s}v_l\right\|^2 \\
&= \frac{R^2}{m}\sum_{l=m+1}^{2m}(\theta_l^{(j)}-\theta_l^{(k)})^2\lambda_l^{2(\mu+s)} \\
&\geq \frac{R^2}{m}\lambda_{2m}^{2(\mu+s)}H(\theta^{(j)},\theta^{(k)}) \\
&\geq \frac{R^2}{8}\lambda_{2m}^{2(\mu+s)} = \frac{R^2}{8}(2m)^{-2(\mu+s)}\log^{-2a(\mu+s)}(2m) \\
&\geq 4A^2n^{-\frac{\mu+s}{\mu+1}}\log^{-2a(\mu+s)}n
\end{aligned}$$

where the last two inequalities follow from (b) provided that $m\leq n^{\frac{1}{2(\mu+1)}}$ for some $A>0$. This verifies (i).

Next, since $Y_i|X_i\sim N(\langle X_i,\beta_{\theta^{(j)}}\rangle,\sigma^2)$ under P_j, we have $P_j << P_0, \forall j=1,\dots,M$ with the log-likelihood ratio

$$\log\frac{\mathrm{d}P_j}{\mathrm{d}P_0}=\frac{1}{\sigma^2}\sum_{i=1}^n\left(Y_i-\langle X_i,\beta_{\theta^{(j)}}\rangle\right)\langle X_i,\beta_{\theta^{(j)}}-\beta_{\theta^{(0)}}\rangle+\frac{1}{2\sigma^2}\sum_{i=1}^n\langle X_i,\beta_{\theta^{(j)}}-\beta_{\theta^{(0)}}\rangle^2.$$

To verify (iii), we compute the Kullback–Leibler divergence:

$$\begin{aligned}
KL(P_j,P_0) &= \int\log\frac{\mathrm{d}P_j}{\mathrm{d}P_0}\mathrm{d}P_j \\
&= \frac{n}{2\sigma^2}\mathbb{E}\langle X_i,\beta_{\theta^{(0)}}-\beta_{\theta^{(j)}}\rangle^2 \\
&= \frac{n}{2\sigma^2}\left\|K^{1/2}(\beta_{\theta^{(0)}}-\beta_{\theta^{(j)}})\right\|^2 \\
&= \frac{n}{2\sigma^2}\left\|Rm^{-1/2}\sum_{l=m+1}^{2m}\theta_l^{(j)}\lambda_l^{\mu+1/2}v_l\right\|^2 \\
&= \frac{nR^2}{2\sigma^2m}\sum_{l=m+1}^{2m}\left(\theta_l^{(j)}\right)^2\lambda_l^{2\mu+1} \\
&\leq \frac{nR^2}{2\sigma^2}\lambda_m^{2\mu+1}=\frac{nR^2}{2\sigma^2}m^{-(2\mu+1)}\left(\log m\right)^{-a(2\mu+1)} \\
&\leq \alpha\frac{m}{8}\log 2=\alpha\log 2^{m/8},
\end{aligned}$$

provided that $m \geq (c_0 n)^{\frac{1}{2(\mu+1)}} (\log m)^{-a\frac{2\mu+1}{2(\mu+1)}}$ with $c_0 = 4R^2/(\sigma^2\alpha\log 2)$ some $\alpha \in (0, 1/8)$.[6] This verifies (iii) in light of (c).

Therefore, by Tsybakov (2009), Theorem 2.5

$$\liminf_{n\to\infty} \inf_{\hat{\beta}} \max_{\theta\in\{\theta^{(0)},\dots,\theta^{(M)}\}} \Pr\left(\left\|K^s(\hat{\beta}-\beta_\theta)\right\| \geq An^{-\frac{\mu+s}{2(\mu+1)}} \log^{-a(\mu+s)} n\right) \geq 1-2\alpha > 0$$

which implies the result in light of the inequality (8). □

The next lemma provides an upper bound for the derivative of the residual polynomial of degree selected by the stopping rule in Assumption 4 with some fixed $\delta \in (0,1)$.

Lemma 2. *Suppose that Assumptions 1, 2, 3, and 4 are satisfied with $\delta \geq 1/n$. Then*

$$|\hat{Q}'_{\hat{m}}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$$

on an event with probability at least $1-\delta$.

Proof of Lemma 2. We have

$$|\hat{Q}'_{\hat{m}}(0)| \leq |\hat{Q}'_{\hat{m}-1}(0)| + |\hat{Q}'_{\hat{m}}(0) - \hat{Q}'_{\hat{m}-1}(0)|,$$

where each of the two terms will be bounded separately.

By the virtue of Assumption 4

$$\begin{aligned}\tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} &\leq \left\|\hat{r} - \hat{K}\hat{\beta}_{\hat{m}-1}\right\| \\ &\leq c\left\{\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + |\hat{Q}'_{\hat{m}-1}(0)|^{-1}\sqrt{\frac{2\mathbf{E}\|X\|^4}{\delta n}} + |\hat{Q}'_{\hat{m}-1}(0)|^{-(\mu+1)}\right\},\end{aligned}$$

where the second line follows by Lemma 1 for some $c > 0$.

[6] To ensure that this constraint holds and that $m \leq n^{\frac{1}{2(\mu+1)}}$, we can take m as a fraction of $n^{\frac{1}{2(\mu+1)}}$.

Therefore,

$$(\tau - c)\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq c\max\left\{|\hat{Q}'_{\hat{m}-1}(0)|^{-1}\sqrt{\frac{1}{\delta n}}, |\hat{Q}'_{\hat{m}-1}(0)|^{-(\mu+1)}\right\}.$$

If the first term inside the maximum is larger, then $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim 1$ while if the second term is larger, then $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$ provided that $\tau > c$. Therefore, we always have $|\hat{Q}'_{\hat{m}-1}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$.

For the second term, by Hanke (1995), Corollary 2.6, for every $1 \leq m \leq n_*$

$$0 \leq \hat{Q}'_{m-1}(0) - \hat{Q}'_m(0) = \frac{[\hat{Q}_{m-1}, \hat{Q}_{m-1}]_0 - [\hat{Q}_m, \hat{Q}_m]_0}{[\hat{Q}^{[2]}_{m-1}, \hat{Q}^{[2]}_{m-1}]_1} \leq \frac{[\hat{Q}_{m-1}, \hat{Q}_{m-1}]_0}{[\hat{Q}^{[2]}_{m-1}, \hat{Q}^{[2]}_{m-1}]_1}, \tag{9}$$

where $(\hat{Q}^{[2]}_l)_{l\geq 0}$ are the polynomials orthogonal with respect to $[.,.]_2$ and constant equal to 1; see equation (5).

Take $a \in (0, \hat{\theta}_{m-1}]$ and let $\hat{K}^+ = \sum_{j=1}^{n_*} \hat{v}_j \otimes \hat{v}_j / \hat{\lambda}_j$ be the generalized inverse of $\hat{K}$. Then

$$\begin{aligned}
\sqrt{[\hat{Q}_{m-1}, \hat{Q}_{m-1}]_0} &= \left\|\hat{Q}_{m-1}(\hat{K})\hat{r}\right\| \leq \left\|\hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r}\right\| \\
&\leq \left\|\Pi_a \hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r}\right\| + \left\|\Pi_a^\perp \sqrt{\hat{K}^+}\hat{K}^{1/2}\hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r}\right\| \\
&\leq \left\|\Pi_a \hat{Q}^{[2]}_{m-1}(\hat{K})\right\|_{\mathrm{op}} \|\Pi_a \hat{r}\| + \left\|\Pi_a^\perp \sqrt{\hat{K}^+}\right\|_{\mathrm{op}} \left\|\hat{K}^{1/2}\hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r}\right\| \\
&\leq \sup_{\lambda\in[0,a]} \left|\hat{Q}^{[2]}_{m-1}(\lambda)\right| \|\Pi_a \hat{r}\| + \sup_{\lambda \geq a} \frac{1}{\sqrt{\lambda}} \left\|\hat{K}^{1/2}\hat{Q}^{[2]}_{m-1}(\hat{K})\hat{r}\right\| \\
&\leq \|\Pi_a \hat{r}\| + \sqrt{[\hat{Q}^{[2]}_{m-1}, \hat{Q}^{[2]}_{m-1}]_1 / a},
\end{aligned}$$

where the second line holds since $\hat{Q}_m$ solves the problem in equation (14) and the last line since $|\hat{Q}^{[2]}_{m-1}(\lambda)| \leq 1, \forall \lambda \in [0, a]$; see the proof of Lemma 6.

Next, under Assumption 3, $\beta = K^\mu w$ with $\|w\| \le R$, so that

$$\begin{aligned}
\|\Pi_a \hat r\| &\le \left\|\Pi_a(\hat r - \hat K\beta)\right\| + \left\|\Pi_a \hat K\beta\right\| \\
&\le \left\|\hat r - \hat K\beta\right\| + \left\|\Pi_a \hat K \hat K^\mu w\right\| + \left\|\Pi_a \hat K(\hat K^\mu - K^\mu)w\right\| \\
&\le \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + \sup_{\lambda\in[0,a]} \lambda^{1+\mu} R + a\left\|\hat K^\mu - K^\mu\right\|_{\mathrm{op}} R \\
&\lesssim \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + a^{\mu+1} + a\left(\frac{1}{\delta n}\right)^{\frac{\mu\wedge 1}{2}},
\end{aligned}$$

where we use the inequality in equation (16) with $\gamma = \delta$. Take $a = (c_1\sigma\sqrt{2\mathbf{E}\|X\|^2/\delta n})^{1/(\mu+1)}$ with a sufficiently small $c_1 > 0$, so that $a \le |\hat Q'_{\hat m-1}(0)|^{-1} \le \hat\theta_{\hat m-1}$, cf. Lemma 6 (iv). Such a constant exists since as we've already shown $|\hat Q'_{\hat m-1}(0)| \lesssim (n\delta)^{\frac{1}{2(\mu+1)}}$. Then for some $c_3 > 0$

$$\begin{aligned}
\sqrt{[\hat Q_{\hat m-1}, \hat Q_{\hat m-1}]_0} &\le c_3\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + \sqrt{[\hat Q^{[2]}_{\hat m-1}, \hat Q^{[2]}_{\hat m-1}]_1/a} \\
&\le \frac{c_3}{\tau}\left\|\hat r - \hat K\hat\beta_{\hat m-1}\right\| + \sqrt{[\hat Q^{[2]}_{\hat m-1}, \hat Q^{[2]}_{\hat m-1}]_1/a} \\
&= \frac{c_3}{\tau}\sqrt{[\hat Q_{\hat m-1}, \hat Q_{\hat m-1}]_0} + \sqrt{[\hat Q^{[2]}_{\hat m-1}, \hat Q^{[2]}_{\hat m-1}]_1/a},
\end{aligned}$$

where we use Assumption 4 and equation (15). If τ is selected so that $\tau > c_3$ in Assumption 4, then

$$[\hat Q_{\hat m-1}, \hat Q_{\hat m-1}]_0 \le \left(\frac{\tau}{\tau - c_3}\right)^2 [\hat Q^{[2]}_{\hat m-1}, \hat Q^{[2]}_{\hat m-1}]_1/a.$$

Plugging this into equation (9) and with our choice of a, we get

$$\left|\hat Q'_m(0) - \hat Q'_{m-1}(0)\right| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}.$$

□

Proof of Theorem 3. Setting $m = \hat m$ and $\gamma = \delta$ in equation (7), under Assumption 4, we obtain

$$\left\|\hat K^s(\hat\beta_{\hat m} - \beta)\right\| \lesssim a^{s-1}(\delta n)^{-1/2} + a^{\mu+s} + a^s(\delta n)^{-\frac{\mu\wedge 1}{2}} + a^s|\hat Q'_{\hat m}(0)|(\delta n)^{-1/2} \tag{10}$$

for every $a \leq |\hat{Q}'_{\hat{m}}(0)|^{-1}$. Now we will choose the truncation level a. Suppose that $s \in [0,1)$. Then the function $a \mapsto a^{s-1}(\delta n)^{-1/2} + a^{\mu+s}$ is minimized at $a^* = \left\{(\delta n)^{1/2}(\mu+s)/(1-s)\right\}^{-\frac{1}{\mu+1}}$. If $a^* \leq |\hat{Q}'_{\hat{m}}(0)|^{-1}$, we shall choose $a = a^*$, in which case since $\delta \geq 1/n$, we obtain

$$\left\|\hat{K}^s(\hat{\beta}_{\hat{m}} - \beta)\right\| \lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}} + (\delta n)^{-\frac{s+\mu+1}{2(\mu+1)}}|\hat{Q}'_{\hat{m}}(0)| \lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}}.$$

On the other hand, if $a^* > |\hat{Q}'_{\hat{m}}(0)|^{-1}$, we shall choose $a = |\hat{Q}'_{\hat{m}}(0)|^{-1}$. Then

$$\begin{aligned}\left\|\hat{K}^s(\hat{\beta}_{\hat{m}} - \beta)\right\| &\lesssim |\hat{Q}'_{\hat{m}}(0)|^{1-s}(\delta n)^{-1/2} + |\hat{Q}'_{\hat{m}}(0)|^{-(\mu+s)} + |\hat{Q}'_{\hat{m}}(0)|^{-s}(\delta n)^{-\frac{\mu\wedge 1}{2}} \\ &\lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}},\end{aligned}$$

where the last line follows from $|\hat{Q}'_{\hat{m}}(0)| \lesssim (\delta n)^{\frac{1}{2(\mu+1)}}$ by Lemma 2, and from $|\hat{Q}'_{\hat{m}}(0)|^{-1} < a^* \lesssim (\delta n)^{-\frac{1}{2(\mu+1)}}$ and $(\delta n)^{-1} \leq 1$.

If $s = 1$, then setting $a = c(\delta n)^{-\frac{1}{2(\mu+1)}}$ in equation (10), for some $c > 0$ such that $a \leq |\hat{Q}'_{\hat{m}}(0)|^{-1}$, cf. Lemma 2, we get

$$\begin{aligned}\left\|\hat{K}^s(\hat{\beta}_{\hat{m}} - \beta)\right\| &\lesssim (\delta n)^{-1/2} + (\delta n)^{-\frac{1+(\mu\wedge 1)(\mu+1)}{2(\mu+1)}} + (\delta n)^{-\frac{\mu+2}{2(\mu+1)}}|\hat{Q}'_{\hat{m}}(0)| \\ &\lesssim (\delta n)^{-1/2},\end{aligned}$$

where the last line follows from Lemma 2 and $(\delta n)^{-1} \leq 1$. Therefore, we've just established for every $s \in [0,1]$

$$\left\|\hat{K}^s(\hat{\beta}_{\hat{m}} - \beta)\right\| \lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}}. \tag{11}$$

This proves the statement of the theorem in the special case when $s = 0$. On the other hand, if $s \in (0,1]$, we have

$$\begin{aligned}\left\|K^s(\hat{\beta}_{\hat{m}} - \beta)\right\| &\leq \left\|\hat{K}^s(\hat{\beta}_{\hat{m}} - \beta)\right\| + \left\|\hat{K}^s - K^s\right\|_{\mathrm{op}}\left\|\hat{\beta}_{\hat{m}} - \beta\right\| \\ &\lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}} + (\delta n)^{-\frac{s\wedge 1}{2}}(\delta n)^{-\frac{\mu}{2(\mu+1)}} \\ &\lesssim (\delta n)^{-\frac{\mu+s}{2(\mu+1)}},\end{aligned}$$

where we use equation (11) and (16), and $(\delta n)^{-1} \leq 1$. □

Proof of Theorem 4. For $1 \leq m \leq n_*$ and $\nu \geq 0$, let $\tilde{P}_m^{(\nu)}$ be a Jacobi polynomial of degree m on $[-1, 1]$, i.e. a polynomial, orthogonal with respect to the weight $\lambda \mapsto (1-\lambda)^\alpha(1+\lambda)^\beta$, where we set $\alpha = -1/2$ and $\beta = 2\nu - 1/2$ with $\nu > 0$. Let $P_m^{(\nu)}(\lambda) = \tilde{P}_m^{(\nu)}(2\lambda/\hat{\lambda}_1 - 1)/\tilde{P}_m^{(\nu)}(-1), \forall \lambda \in [0, \hat{\lambda}_1]$ be a shifted Jacobi polynomial, normalized so that $P_m^{(\nu)}(0) = 1$. By Engl et al. (1996), Appendix A.2, p.294, there exists $c_\nu > 0$ such that

$$|P_m^{(\nu)}(\lambda)| \leq c_\nu(1 + m^2\lambda)^{-\nu}, \qquad \forall \lambda \in [0, \hat{\lambda}_1],\ m \geq 0. \tag{12}$$

Recall also that under Assumption 3, we have $\beta = K^\mu w$ with $\|w\| \leq R$. Then

$$\begin{aligned}
\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| &= \left\|\hat{Q}_m(\hat{K})\hat{r}\right\| \\
&\leq \left\|P_m^{(\mu+1)}(\hat{K})\hat{r}\right\| \\
&\leq \left\|P_m^{(\mu+1)}(\hat{K})(\hat{r} - \hat{K}\beta)\right\| + \left\|P_m^{(\mu+1)}(\hat{K})\hat{K}\beta\right\| \\
&\leq \left\|P_m^{(\mu+1)}(\hat{K})\right\|_{\text{op}}\left\|\hat{r} - \hat{K}\beta\right\| + \left\|P_m^{(\mu+1)}(\hat{K})\hat{K}\hat{K}^\mu w\right\| \\
&\qquad + \left\|P_m^{(\mu+1)}(\hat{K})\hat{K}(\hat{K}^\mu - K^\mu)w\right\| \\
&\lesssim \sup_{\lambda \in [0,\hat{\lambda}_1]} |P_m^{(\mu+1)}(\lambda)| \left\|\hat{r} - \hat{K}\beta\right\| + \sup_{\lambda \in [0,\hat{\lambda}_1]} \left|\lambda^{\mu+1}P_m^{(\mu+1)}(\lambda)\right| \\
&\qquad + \sup_{\lambda \in [0,\hat{\lambda}_1]} |\lambda P_m^{(\mu+1)}(\lambda)| \left\|\hat{K}^\mu - K^\mu\right\|_{\text{op}} \\
&\lesssim \left\|\hat{r} - \hat{K}\beta\right\| + m^{-2(\mu+1)} + \hat{\lambda}_1 \left\|\hat{K}^\mu - K^\mu\right\|_{\text{op}} \\
&\lesssim \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} + m^{-2(\mu+1)} + \|\hat{K}\|_{\text{op}}\left(\frac{2\mathbf{E}\|X\|^4}{\gamma n}\right)^{\frac{\mu\wedge 1}{2}},
\end{aligned}$$

where the second line follows since $\hat{Q}_m$ minimizes the problem in equation (14); the sixth from the inequality (12); and the last by Lemma 4 and the inequality (16) with probability at least $1 - \gamma$ for every $\gamma \in (0, 1)$.

Therefore, if $\mu \geq 1$, under Assumption 1 by Lemma 4 and the inequality (16), we obtain

$$\left\| \hat{r} - \hat{K}\hat{\beta}_m \right\| \leq c \left\{ \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + m^{-2(\mu+1)} \right\}$$

for some $c > 0$ and $\delta \geq 1/n$ on the event with probability at least $1 - \delta$. According to the stopping rule in the Assumption 4, we also know that

$$\tau\sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq \left\| \hat{r} - \hat{K}\hat{\beta}_{\hat{m}-1} \right\|$$

Therefore,

$$(\tau - c)\sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \lesssim (\hat{m} - 1)^{-2(\mu+1)},$$

and whence $\hat{m} \lesssim (\delta n)^{\frac{1}{4(\mu+1)}}$, provided that $\tau > c$. □

Proof of Theorem 5. For some integers $m \geq k \geq 0$, put

$$G_m(\lambda) := \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\hat{\lambda}_j} \right) P_{m-k}^{(\mu+1)} \left(\frac{\lambda}{\hat{\lambda}_{k+1}} \right)$$

where $P_m^{(\nu)}(\lambda) = \tilde{P}_m^{(\nu)}(2\lambda/\hat{\lambda}_{k+1} - 1)/\tilde{P}_m^{(\nu)}(-1)$ is a shifted and normalized Jacobi polynomial on $[0, \hat{\lambda}_{k+1}]$, defined to be zero outside of this interval; see the proof of Theorem 4. Then G_m is an m^{th} degree polynomial with $G_m(0) = 1$ and

$$\sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} \left| \lambda^{\mu+1} G_m(\lambda) \right| \leq \sup_{\lambda \in [0, \hat{\lambda}_{k+1}]} \left| \lambda^{\mu+1} P_{m-k}^{(\mu+1)} \left(\frac{\lambda}{\hat{\lambda}_{k+1}} \right) \right| \leq \hat{\lambda}_{k+1}^{\mu+1} c_\mu (m - k)^{-2(\mu+1)},$$

where we use the inequality (12). Then since under Assumption 3, $\beta = K^\mu w$ with $\|w\| \leq R$, we have with probability at least $1-\gamma$ for every $\gamma \in (0,1)$

$$\begin{aligned}
\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| &= \left\|\hat{Q}_m(\hat{K})\hat{r}\right\| \leq \left\|G_m(\hat{K})\hat{r}\right\| \\
&\leq \left\|G_m(\hat{K})(\hat{r} - \hat{K}\beta)\right\| + \left\|G_m(\hat{K})\hat{K}K^{\mu}w\right\| + \left\|G_m(\hat{K})\hat{K}(\hat{K}^{\mu} - K^{\mu})w\right\| \\
&\leq \sup_{\lambda \in [0,\hat{\lambda}_{k+1}]} |G_m(\lambda)| \left\|\hat{r} - \hat{K}\beta\right\| + R \sup_{\lambda \in [0,\hat{\lambda}_{k+1}]} |\lambda^{\mu+1}G_m(\lambda)| \\
&\qquad + R \sup_{\lambda \in [0,\hat{\lambda}_{k+1}]} |\lambda G_m(\lambda)| \left\|\hat{K}^{\mu} - K^{\mu}\right\| \\
&\lesssim \sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} + \hat{\lambda}_{k+1}^{\mu+1}(m-k)^{-2(\mu+1)} + \|\hat{K}\|_{\mathrm{op}} \left(\frac{2\mathbf{E}\|X\|^2}{\gamma n}\right)^{\frac{\mu \wedge 1}{2}}
\end{aligned}$$

where the first inequality follows since $\hat{Q}_m$ solves the problem in equation (14) and for the last inequality, we use $|G_m(\lambda)| \leq |P_{m-k}^{(\mu+1)}(\lambda/\hat{\lambda}_{k+1})| \lesssim 1$; see equation (12). Recall that $\|\hat{K}\|_{\mathrm{op}} \lesssim 1$ on an event with probability at least $1-\gamma$ for $\gamma \geq 1/n$; see Lemma 4. Therefore, since $\mu \geq 1$, we obtain

$$\left\|\hat{r} - \hat{K}\hat{\beta}_m\right\| \leq c\left\{\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} + \hat{\lambda}_{k+1}^{\mu+1}(\hat{m}-k)^{-2(\mu+1)}\right\}$$

for some $c > 0$ and $\delta \geq 1/n$ on an event with probability at least $1-\delta$.

According to the stopping rule in the Assumption 4, we also know that for some $\delta \in (0,1)$

$$\tau\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq \left\|\hat{r} - \hat{K}\hat{\beta}_{\hat{m}-1}\right\|.$$

Therefore, if $\tau > c$, we obtain

$$(\tau - c)\sigma\sqrt{\frac{2\mathbf{E}\|X\|^2}{\delta n}} \leq c\hat{\lambda}_{k+1}^{\mu+1}(\hat{m}-k-1)^{-2(\mu+1)}$$

which implies that

$$\begin{aligned} \hat{m} - k - 1 &\lesssim \hat{\lambda}_{k+1}^{1/2} (\delta n)^{\frac{1}{4(\mu+1)}} \\ &\leq (\delta n)^{\frac{1}{4(\mu+1)}} \left\{ \lambda_{k+1}^{1/2} + \left\| \hat{K}^{1/2} - K^{1/2} \right\|_{\mathrm{op}} \right\} \\ &\lesssim (\delta n)^{\frac{1}{4(\mu+1)}} \left\{ \lambda_{k+1}^{1/2} + (\delta n)^{-1/4} \right\}, \end{aligned} \tag{13}$$

where we use Weyl's inequality and equation (16).

Case (i): if $\lambda_k = O(k^{-2\kappa})$ with $\kappa > 0$, we can take $k \sim \hat{m}/2$. In this case equation (13) implies

$$\hat{m} \lesssim (\delta n)^{\frac{1}{4(\mu+1)}} \hat{m}^{-\kappa} + (\delta n)^{-\frac{\mu}{4(\mu+1)}}.$$

If the second term in this upper bound dominates the first one, then $\hat{m} \lesssim (\delta n)^{-\frac{\mu}{4(\mu+1)}} = o(1)$, which is a contradiction. Therefore, $\hat{m} \lesssim (\delta n)^{\frac{1}{4(\kappa+1)(\mu+1)}}$.

Case (ii): if $\lambda_j = O(q^j)$ with $q \in (0,1)$, we can take $k = \hat{m} - 2$. In this case equation (13) implies

$$1 \lesssim (\delta n)^{\frac{1}{4(\mu+1)}} q^{(\hat{m}-1)/2} + (\delta n)^{-\frac{\mu}{4(\mu+1)}}.$$

If the second term in this upper bound dominates the first one, then $1 \lesssim (\delta n)^{-\frac{\mu}{4(\mu+1)}} = o(1)$, which is a contradiction. Therefore, $\hat{m} \lesssim 1 + \log(\delta n)$ since $\log q < 0$. □

Proof of Theorem 6. Recall that $\hat{Q}_m(\lambda) = 1 - \lambda \hat{P}_m(\lambda)$. Let $(\hat{v}_j)_{j=1}^{\infty}$ be a basis of $\mathbb{H}$, where the first n_* terms correspond to the eigenbasis of $\hat{K}$. Then $\hat{r} = \sum_{j=1}^{\infty} \langle \hat{r}, \hat{v}_j \rangle \hat{v}_j$ and for every

$m \leq n_*$

$$
\begin{aligned}
\left\|\hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PLS}}\right\|^2 &= \left\|\hat{Q}_m(\hat{K})\hat{r}\right\|^2 \\
&= \sum_{j=1}^{n_*} \hat{Q}_m^2(\hat{\lambda}_j)\langle \hat{r}, \hat{v}_j\rangle^2 \\
&\leq \sum_{j=1}^{n_*} \hat{Q}_m(\hat{\lambda}_j)\langle \hat{r}, \hat{v}_j\rangle^2 \\
&= \sum_{j=1}^{n_*} \sum_{n_* \geq j_1 > \cdots > j_m \geq 1} \hat{w}_{j_1,\dots,j_m} \prod_{k=1}^{m}\left(1 - \frac{\hat{\lambda}_j}{\hat{\lambda}_{j_k}}\right)\langle \hat{r}, \hat{v}_j\rangle^2 \\
&\leq \sum_{j=1}^{n_*} \max_{n_* \geq j_1 > \cdots > j_m \geq 1} \prod_{k=1}^{m}\left(1 - \frac{\hat{\lambda}_j}{\hat{\lambda}_{j_k}}\right)\langle \hat{r}, \hat{v}_j\rangle^2 \\
&= \sum_{j=m+1}^{n_*} \prod_{k=1}^{m}\left(1 - \frac{\hat{\lambda}_j}{\hat{\lambda}_k}\right)\langle \hat{r}, \hat{v}_j\rangle^2 \leq \sum_{j=m+1}^{\infty} \langle \hat{r}, \hat{v}_j\rangle^2 \\
&= \left\|\sum_{j=1}^{\infty}\langle \hat{r}, \hat{v}_j\rangle \hat{v}_j - \sum_{j=1}^{m}\frac{1}{\hat{\lambda}_j}\langle \hat{r}, \hat{v}_j\rangle \hat{K}\hat{v}_j\right\|^2 \\
&= \left\|\hat{r} - \hat{K}\hat{\beta}_m^{\mathrm{PCA}}\right\|^2,
\end{aligned}
$$

where the second and last lines follow by Parseval's identity; the third by Blazère et al. (2014), Lemma 3.6, (2); the fourth and fifth by Lemma 6, (vii); and the sixth since the maximum is $= 0$ for all $j \leq m$ and is ≤ 1 if $j > m$ as $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_{n_*}$.

For the second part, put $Q_m(\lambda) = 1 - \lambda P_m(\lambda)$, where $\beta_m^{\mathrm{PLS}} = P_m(K)r$ solves the population counterpart to the problem in equation (3). Similarly, since $\beta = \sum_{j=1}^{\infty}\langle \beta, v_j\rangle v_j$

and $K\beta = r$, we have

$$\begin{aligned}
\left\|K^s(\beta_m^{\mathrm{PLS}} - \beta)\right\|^2 &= \|K^s Q_m(K)\beta\|^2 \\
&= \sum_{j=1}^{\infty} \lambda_j^{2s} Q_m^2(\lambda_j)\langle\beta, v_j\rangle^2 \\
&\leq \sum_{j=1}^{\infty} \lambda_j^{2s} Q_m(\lambda_j)\langle\beta, v_j\rangle^2 \\
&= \sum_{j=1}^{\infty} \lambda_j^{2s} \sum_{j_1 > \cdots > j_m \geq 1} w_{j_1,\ldots,j_m} \prod_{k=1}^{m} \left(1 - \frac{\lambda_j}{\lambda_{j_k}}\right) \langle\beta, v_j\rangle^2 \\
&\leq \sum_{j=m+1}^{\infty} \lambda_j^{2s} \langle\beta, v_j\rangle^2 \\
&= \left\|K^s \left(\sum_{j=1}^{\infty} \langle\beta, v_j\rangle v_j - \sum_{j=1}^{m} \frac{1}{\lambda_j} \langle K\beta, v_j\rangle v_j\right)\right\|^2 \\
&= \left\|K^s(\beta - \beta_m^{\mathrm{PCA}})\right\|^2 .
\end{aligned}$$

□

B Supplementary Results

In this appendix, we collect several supplementary results from various references. The following proposition states that the PLS estimator $\hat{\beta}_m$ is unique for every $m \leq n_*$ and that the tuning parameter selected in Assumption 4 does not exceed the number of unique non-zero eigenvalues, n_*; see also Blanchard & Krämer (2016) for a kernel regression model setting.

Proposition 1. *The solution in equation (3) is unique for every $m \leq n_*$. Moreover, $\hat{m} \leq n_*$.*

Proof of Proposition 1. Let $\mathcal{P}_m$ be the space of real polynomials of degree at most m and let $\mathcal{P}_m^0$ be its subspace of polynomials with constant equal to one. The PLS problem in equation (3) amounts to fitting a polynomial of degree $m-1$ solving

$$\hat{P}_m \in \underset{\phi \in \mathcal{P}_{m-1}}{\arg\min} \left\| \left[I - \hat{K}\phi(\hat{K}) \right] \hat{r} \right\|^2$$

or equivalently a residual polynomial $\hat{Q}_m(\lambda) = 1 - \lambda \hat{P}_m(\lambda)$ solving

$$\hat{Q}_m \in \underset{\phi \in \mathcal{P}_m^0}{\arg\min} \left\| \phi(\hat{K})\hat{r} \right\|^2 . \tag{14}$$

By Parseval's identity, for every $\phi : [0, \hat{\lambda}_1] \to \mathbb{R}$, the objective function can be written as

$$\left\| \phi(\hat{K})\hat{r} \right\|^2 = \sum_{j=1}^{n_*} \phi(\hat{\lambda}_j)^2 \langle \hat{r}, \hat{v}_j \rangle^2 = [\phi, \phi]_0, \tag{15}$$

where $[.,.]_0$ is defined in equation (5). Therefore $\hat{Q}_m$ minimizes $\phi \mapsto [\phi, \phi]_0$ on $\mathcal{P}_m^0$. It is easy to see that $[.,.]_0$ is an inner product for every $m \leq n_* - 1$. Therefore, $\hat{Q}_m$ is the unique projection of zero on a closed subspace $\mathcal{P}_m^0 \subset \mathcal{P}_m$ with respect to $[.,.]_0$. For $m = n_*$, $[.,.]_0$ is not an inner product because we can take an n_*-degree polynomial $\phi \neq 0$ with roots equal to the distinct n_* eigenvalues of $\hat{K}$, so that $[\phi, \phi]_0 = 0$. However, such a polynomial is unique. Therefore, $\hat{P}_m$ and $\hat{\beta}_m$ are unique for every $m \leq n_*$. This also shows that the

PLS objective function is minimized to zero for $m \geq n_*$, so that the tuning parameter in Assumption 4 satisfies $\hat{m} \leq n_*$. □

We will need the following tail inequality in Hilbert spaces.

Lemma 3. *Let $(\xi_i)_{i=1}^n$ be i.i.d. random variables in a Hilbert space $(\mathbb{H}, \langle ., . \rangle)$ with the induced norm $\|.\|$. Suppose that $\mathbf{E}\xi_i = 0$ and $\mathbf{E}\|\xi_i\|^2 < \infty$. Then for every $\gamma \in (0,1)$*

$$\Pr\left(\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\| \leq \sqrt{\frac{\mathbf{E}\|\xi_i\|^2}{\gamma n}} \right) \geq 1 - \gamma.$$

Proof of Lemma 3. By Markov's inequality, $\forall u > 0$

$$\begin{aligned} \Pr\left(\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\| > u \right) &\leq u^{-2} \mathbf{E} \left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\|^2 \\ &= \frac{1}{u^2 n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{E} \langle \xi_i, \xi_j \rangle \\ &= \frac{1}{u^2 n} \mathbf{E} \|\xi_i\|^2, \end{aligned}$$

where the last two lines follow under the i.i.d. hypothesis. Setting $\gamma = \mathbf{E}\|\xi_i\|^2/(nu^2)$ and solving for u gives the result. □

Lemma 3 allows us to control the tail probabilities for the PLS residual as well as the covariance operator errors on an event with probability at least $1 - \gamma$.

Lemma 4. *Suppose that Assumption 1 is satisfied. Then for every $\gamma \in (0,1)$*

$$\left\| \hat{r} - \hat{K}\beta \right\| \leq \sigma \sqrt{\frac{2\mathbf{E}\|X\|^2}{\gamma n}} \qquad \textit{and} \qquad \left\| \hat{K} - K \right\|_{\mathrm{HS}} \leq \sqrt{\frac{2\mathbf{E}\|X\|^4}{\gamma n}}$$

with probability at least $1 - \gamma$, where $\|.\|_{\mathrm{HS}}$ is the Hilbert-Schmidt norm.

Proof of Lemma 4. We will apply Lemma 3. First, we note that

$$\left\| \hat{r} - \hat{K}\beta \right\| = \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|,$$

where $\mathbf{E}\|\varepsilon_i X_i\|^2 \leq \sigma^2 \mathbf{E}\|X_i\|^2 < \infty$ under Assumption 1. Then by Lemma 3 with probability at least $1-\gamma/2$, we have $\|\hat{r} - \hat{K}\beta\| \leq \sigma\sqrt{2\mathbf{E}\|X\|^2/\gamma n}$. Second, the space of Hilbert-Schmidt operators is a Hilbert space and

$$\left\|\hat{K} - K\right\|_{\mathrm{HS}} = \left\|\frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i - \mathbf{E}[X_i \otimes X_i]\right\|_{\mathrm{HS}},$$

where $\mathbf{E}\|X_i \otimes X_i - \mathbf{E}[X_i \otimes X_i]\|_{\mathrm{HS}}^2 \leq \mathbf{E}\|X_i \otimes X_i\|_{\mathrm{HS}}^2 = \mathbf{E}\|X_i\|^4$. Then by Lemma 3 with probability at least $1-\gamma/2$, we have $\|\hat{K} - K\|_{\mathrm{HS}} \leq \sqrt{2\mathbf{E}\|X\|^4/\gamma n}$. The result follows by the union bound. □

We will also use the following two inequalities known in the perturbation theory.

Lemma 5. *Let $A : \mathbb{H} \to \mathbb{H}$ and $B : \mathbb{H} \to \mathbb{H}$ be two self-adjoint Hilbert-Schmidt operators. Then*

$$\|A^\mu - B^\mu\|_{\mathrm{op}} \leq c_\mu \|A - B\|_{\mathrm{op}}^\mu, \qquad 0 < \mu < 1$$

and

$$\|A^\mu - B^\mu\|_{\mathrm{HS}} \leq \mu\nu^{\mu-1}\|A - B\|_{\mathrm{HS}}, \qquad \mu \geq 1,$$

where $\nu = \|A\|_{\mathrm{op}} \vee \|B\|_{\mathrm{op}}$.

Proof. See Aleksandrov & Peller (2016), Theorem 1.7.2 for the first inequality. The second inequality follows from Aleksandrov & Peller (2016), Theorem 3.5.1. □

As an immediate consequence of Lemmas 4 and 5, since $\|.\|_{\mathrm{op}} \leq \|.\|_{\mathrm{HS}}$, for every $\gamma \in (0,1)$, we have

$$\left\|\hat{K}^\mu - K^\mu\right\|_{\mathrm{op}} \leq (c_\mu \mathbf{1}_{\mu \leq 1} + \mu\nu^{\mu-1}\mathbf{1}_{\mu > 1})\left(\frac{2\mathbf{E}\|X\|^4}{\gamma n}\right)^{\frac{\mu \wedge 1}{2}} \tag{16}$$

on an event that holds with probability at least $1-\gamma$, where $\nu = \|\hat{K}\|_{\mathrm{op}} \vee \|K\|_{\mathrm{op}}$.

The following Lemma presents some useful results on the residual polynomials $\hat{Q}_m(\lambda) = 1 - \lambda\hat{P}_m(\lambda)$; see Engl et al. (1996) and Hanke (1995). For the completeness of the presentation, we sketch proofs for the key results and refer to the aforementioned monographs for others.

Lemma 6. *Let $m \leq n_*$ be a positive integer. Then*

(i) $\hat{Q}_m$ *has* m *distinct positive real roots, denoted* $\hat{\theta}_1 > \hat{\theta}_2 > ... > \hat{\theta}_m > 0$.

(ii) $\hat{Q}_m$ *is positive, decreasing, and convex on* $[0, \hat{\theta}_m]$.

(iii) $(\hat{Q}_l)_{l=0}^{n_*}$ *are orthogonal with respect to* $[.,.]_1$.

(iv) $|\hat{Q}_m'(0)|^{-1} \leq \hat{\theta}_m$.

(v) $\hat{Q}_m(\lambda) = \hat{Q}_{m-1}(\lambda)(1 - \lambda/\hat{\theta}_m)$.

(vi) $\sup_{\lambda \in [0,\hat{\theta}_m]} \lambda^\delta \hat{Q}_m(\lambda)\sqrt{\hat{\theta}_m/(\hat{\theta}_m - \lambda)} \leq (2\delta)^\delta |\hat{Q}_m'(0)|^{-\delta}$ *for every* $\delta \geq 0$ *with* $0^0 := 1$.

(vii) we have

$$\hat{Q}_m(\lambda) = \sum_{n_* \geq j_1 > \cdots > j_m \geq 1} \hat{w}_{j_1,\ldots,j_m} \prod_{k=1}^{m} \left(1 - \frac{\lambda}{\hat{\lambda}_{j_k}}\right)$$

for some $\hat{w}_{j_1,\ldots,j_m} \in (0,1]$ *satisfying* $\sum_{n_* \geq j_1 > \cdots > j_m \geq 1} \hat{w}_{j_1,\ldots,j_m} = 1$.

Proof of Lemma 6. (i) is known in the theory of orthogonal polynomials; see Engl et al. (1996), Appendix A.2. For (iii) and (vi), see Engl et al. (1996), Corollary 7.4, and equation (7.8).

Note that since $\hat{Q}_m(0) = 1$, we can write

$$\hat{Q}_m(\lambda) = \prod_{j=1}^{m} \left(1 - \frac{\lambda}{\hat{\theta}_j}\right).$$

This equation implies (v). Moreover, for all $\lambda \in [0, \hat{\theta}_m]$, we have $\hat{Q}_m(\lambda) \geq 0$ and by (i)

$$\hat{Q}_m'(\lambda) = -\sum_{k=1}^{m} \frac{1}{\hat{\theta}_k} \prod_{j \neq k} \left(1 - \frac{\lambda}{\hat{\theta}_j}\right) \leq 0.$$

We also have $\hat{Q}_m''(\lambda) \geq 0$ for all $\lambda \in [0, \hat{\theta}_m]$ which proves (ii).

(iv) follows from (i) and

$$|\hat{Q}'_m(0)| = \sum_{k=1}^{m} \frac{1}{\hat{\theta}_k} \geq \frac{1}{\hat{\theta}_m}.$$

Lastly, the proof of (vii) is similar to Blazere et al. (2014), Theorem 4.1. □

References

Aleksandrov, A. B., & Peller, V. V. (2016). Operator lipschitz functions. *Russian Mathematical Surveys*, *71*(4), 605–702.

Babii, A., & Florens, J.-P. (2017). Is completeness necessary? estimation in nonidentified linear models. *arXiv preprint arXiv:1709.03473*.

Blanchard, G., Hoffmann, M., & Reiß, M. (2018). Optimal adaptation for early stopping in statistical inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, *6*(3), 1043–1075.

Blanchard, G., & Krämer, N. (2016). Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, *14*(06), 763–794.

Blanchard, G., & Mathé, P. (2012). Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse problems*, *28*(11), 115011.

Blazere, M., Gamboa, F., & Loubes, J.-M. (2014). Pls: a new statistical insight through the prism of orthogonal polynomials. *arXiv preprint arXiv:1405.5900*.

Blazère, M., Gamboa, F., & Loubes, J.-M. (2014). A unified framework to study the properties of the pls vector of regression coefficients. In *International conference on partial least squares and related methods* (pp. 227–237).

Cai, T. T., & Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, *34*(5), 2159–2179.

Cai, T. T., & Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, *107*(499), 1201–1216.

Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, *45*(1), 11–22.

Cardot, H., Ferraty, F., & Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, *13*(3), 571–591.

Cardot, H., & Johannes, J. (2010). Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, *101*(2), 395–408.

Carrasco, M., Florens, J.-P., & Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, *6*, 5633–5751.

Carrasco, M., & Rossi, B. (2016). In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, *34*(3), 313–338.

Cavalier, L. (2011). Inverse problems in statistics. In *Inverse problems and high-dimensional estimation: Stats in the château summer school, august 31-september 4, 2009* (pp. 3–96). Springer.

Comte, F., & Johannes, J. (2012). Adaptive functional linear regression. *The Annals of Statistics*, *40*(6), 2765–2797.

Crambes, C., Kneip, A., & Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, *37*(1), 35–72.

Delaigle, A., & Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, *40*(1), 322–352.

Engl, H. W., Hanke, M., & Neubauer, A. (1996). *Regularization of inverse problems* (Vol. 375). Springer Science & Business Media.

Hall, P., & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, *35*(1), 70–91.

Hanke, M. (1995). *Conjugate gradient type methods for linear ill-posed problems.* Pitman Research Notes in Mathematics Series.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, *17*(2), 581–607.

Hestenes, M. R., Stiefel, E., et al. (1952). Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, *49*(6), 409–436.

Hoffmann, M., & Reiss, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*(1), 310–336.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*(3), 300–303.

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, *186*(2), 294–316.

Klemelä, J., & Mammen, E. (2010). Empirical risk minimization in inverse problems. *The Annals of Statistics*, *38*(1), 482–511.

Kress, R. (1999). *Linear integral equations* (Vol. 82). Springer Science & Business Media.

Nemirovski, A. S. (1986). On regularizing properties of the conjugate gradient method for ill-posed problems (in russian). *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, *26*(3), 332–347.

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization.* Springer.

Phatak, A., & de Hoog, F. (2002). Exploiting the connection between pls, lanczos methods and conjugate gradients: alternative proofs of some properties of pls. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *16*(7), 361–367.

Preda, C., & Saporta, G. (2005). Clusterwise pls regression on a stochastic process. *Computational Statistics & Data Analysis*, *49*(1), 99–108.

Reiss, P. T., & Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, *102*(479), 984–996.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Science & Business Media.

Wold, S., Ruhe, A., Wold, H., & Dunn, W., III. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743.